

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property
Organization
International Bureau



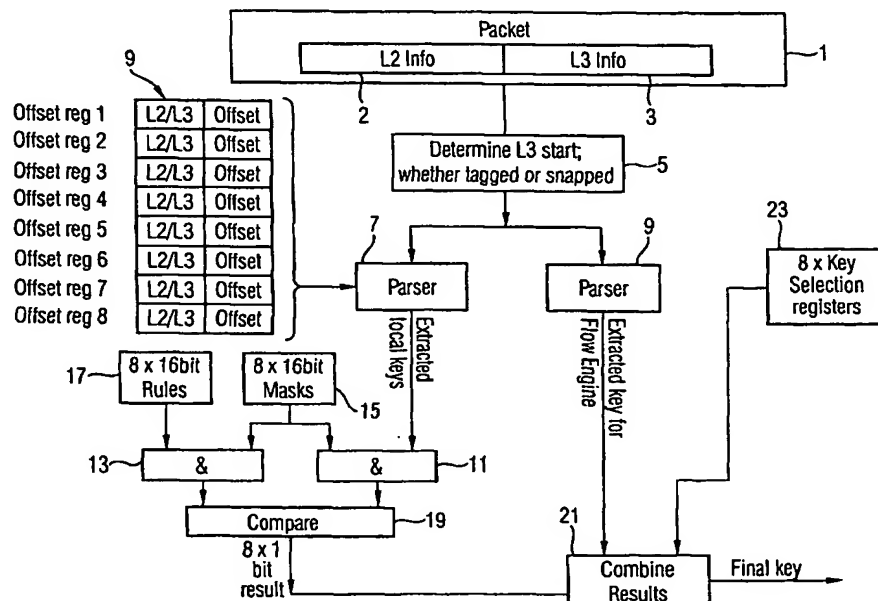
(43) International Publication Date
18 March 2004 (18.03.2004)

PCT

(10) International Publication Number
WO 2004/023762 A1

- (51) International Patent Classification⁷: **H04L 29/06**, 12/46
- (21) International Application Number: **PCT/SG2002/000210**
- (22) International Filing Date: **6 September 2002 (06.09.2002)**
- (25) Filing Language: **English**
- (26) Publication Language: **English**
- (71) Applicant (for all designated States except US): **INFINEON TECHNOLOGIES AG [DE/DE]; St.-Martin-Strasse 53, 81669 Munich (DE).**
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): **MISHRA, Shridhar, Mubaraq [IN/US]; 1325A Spruce Street, Berkeley, CA-94709 (US). HU, Chunfeng [CN/SG]; Blk 105 Bedok Reservoir Rd #12-398, Singapore 470105 (SG).**
- (74) Agent: **WATKIN, Timothy, Lawrence, Harvey; Lloyd Wise, Tanjong Pagar, P.O. Box 636, Singapore 910816 (SG).**
- (81) Designated States (national): **AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.**
- (84) Designated States (regional): **ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).**
- Declaration under Rule 4.17:**
— of inventorship (Rule 4.17(iv)) for US only
- Published:**
— with international search report
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: **A PARSER FOR PARSING DATA PACKETS**



(57) Abstract: A parser system is arranged to receive a data stream (1) having interleaved sections derived from a plurality of different packets, and to extract data from each section as it arrives. The parser system has a scanning section which receives information about each of the sections of data defining which packet it relates to, and employs this information and the properties of the data stream, to identify the locations of layer (2), layer (3) and layer (4) data. This information is passed to parser units (7), (9) which extract data based on this data and also offsets. The offsets for the parser (7) are stored in user-programmable registers (9).

A parser for parsing data packets

Related applications

The present rule is a group of five patent applications having the same priority date. Application PCT/SG02/-----relates to an switch having an ingress port
5 which is configurable to act either as eight FE (fast Ethernet) ports or as a GE (gigabit Ethernet port). The present application relates to a parser suitable for use in such as switch. Application PCT/SG02/----- relates to a flow engine suitable for using the output of the parser to make a comparison with rules. Application PCT/SG02/----- relates to monitoring bandwidth consumption
10 using the results of a comparison of rules with packets. Application PCT/SG02/----- relates to a combination of switches arranged as a stack. The respective subjects of the each of the group of applications have applications other than in combination with the technology described in the other four applications, but the disclosure of the other applications of the group is
15 incorporated by reference.

Field of the invention

The present invention relates to a parser system for parsing multiple data packets to extract information from them. In a particular example, the parser system may be employed in a switch such as an Ethernet switch to parse
20 received data packets and thereby obtain information which is required for processing the packet by the queue management system and switching fabric.

Background of Invention

Recent advances in Internet technology have changed the way we exchange
25 information. Ubiquitous use of the Internet has led to the idea of convergence. The different types of data (e.g. video, sounds, pictures and text) must traverse the same network, and this has given rise to a plethora of protocols

which aim to transmit real time and data traffic on a single network with quality of service support.

Chief among these protocols are DiffServ, IntServ and MPLS, which each require packet classification (i.e. determination of the packet's type) in real time when it is received. The first step in this classification is to extract
5 relevant bytes from a packet, "parsing". This is described in Chapter 1 of "Computer Networks", Andrew S Tanenbaum, Prentice Hall 2nd Ed, 1988.

The parsing operation includes determining whether the packet includes tags. For example, conventionally Ethernet packets may include a VLAN (virtual
10 local area network) tag – i.e. a tag which indicates a VLAN associated with the packet. A VLAN tag is conventionally 4-bytes inserted into the Ethernet frame between the Source MAC Address field and the Length/Type field. The first 2 bytes of the VLAN tag are always set to a value of 0x8100, while the second two bytes are control information (user priority field, canonical format
15 indicator and VLAN identifier).

Another type of tag is defined by the SNAP protocol (subnetwork access protocol), which was introduced to allow older frames and protocols to be encapsulated in a Type 1 LLC header so making any protocol 'pseudo-IEEE
20 compliant'. The SNAP tag (or "snap encapsulation") is placed directly after the standard length/type field of the Ethernet packet (which always takes a value less than or equal to 1500), and has AA-AA as its first two bytes. A packet containing a SNAP tag is called SNAPped.

25 Generally, an Ethernet packet is made up of levels of nested data, known as layers. Data which is interpreted directly by a machine is called "layer 1", or physical layer, data. "Layer 2", or data link layer, data is LAN (local area network) data, such as MAC (media access control) data uniquely identifying an adapter on the LAN. Within the "layer 2" packet may be "layer 3", or

network layer, data defining among other things the IP source address and destination address of the packet. Within the layer 3 packet may be "layer 4" data, or transport layer data, e.g. TCP (transmission control protocol) data.

- 5 In view of the great variety of protocols which may be encountered, it would be useful to provide a parsing technique which is highly flexible.

Additionally, there are a variety of circumstances in which it would be useful to parse a plurality of concurrently received data packets. Such circumstances
10 are not limited to Ethernet applications, but, taking Ethernet applications as an example, in a co-pending pending application referred to above, the present inventors propose a configurable Ethernet switch which can function both as a Fast Ethernet and as a Gigabit Ethernet switch, in order to facilitate the transition from FE to GE Ethernet. A data port can be operated as eight FE
15 MAC interfaces or alternatively as a single GE MAC interface. In the former case, it may happen that the eight FE interfaces receive packets at the same time. It would be possible to provide sufficiently large buffers at the input port that all of these packets are completely received before the processing of any one of the packets begins, but this increases the cost of the buffers required.
20 It would instead be useful to be able to process all of the packets concurrently ("on-the-fly") as they are received to reduce the buffering requirement.

Summary of the Invention

A first aspect of the invention proposes in general terms that a parser system
25 is arranged to receive a data stream having interleaved sections derived from a plurality of different packets, and to extract data from each section as it arrives. The parser system receives information about each of the sections of data defining which packet it relates to, and employs this information to identify data to be extracted from the data stream.

A second aspect of the invention relates to a parser system having a number of programmable registers which store data for the parser, the parser system receiving a data stream and extracting data from it based on offset information stored in the programmable registers.

- 5 In either aspect of the invention, the parser system preferably includes a scanning section which identifies the location of major structural features of data in the data stream (e.g. a location where one of the layers of data commences), and at least one parser unit which uses the output of the scanning section and offset information (at least partly from the programmable registers in the case of the second aspect of the invention) to extract the data.
10 The offset information identifies the offset of the data to be extracted from the location of the structural features.

The scanning section uses the received information about the data stream, and also examines the data itself to identify characteristics of packets in the data stream. For example, in addition to determining the location of the start
15 of any one or more of layer 2, layer 3 and/or layer 4 data in the packets, it may further be able to identify if the packet is VLAD tagged and/or SNAPed.

There may be two parser units, one of which extracts data according to predefined offsets and the structural data from the scanning section, and the other extracting data according to the structural data from the scanning
20 section and offsets defined by the programmable registers.

Brief Description of The Figures

Preferred features of the invention will now be described, for the sake of illustration only, with reference to the following figures in which:

- 25 Fig. 1 shows schematically the operation of a parser system which is an embodiment of the invention;

Fig. 2 shows schematically how key extraction is performed by the second parser and combiner of the embodiment of Fig. 1;

Fig. 3 shows the structure of 4 types of packets to be parsed by the embodiment; and

5 Fig. 4 shows the parser system of Fig. 1 as a circuit diagram.

Detailed Description of the embodiments

Referring firstly to Fig. 1, the operation of the embodiment is shown
10 schematically.

The embodiment processes a data stream having packets 1 containing layer 2 data 2 (starting at bit 0) and layer 3 data 3. In fact the packet will also contain layer 4 data, but for the purposes of this embodiment this may be treated
15 simply as part of the level 3 data. The position of the start of the layer 4 data is given by a field which is part of the layer 3 data.

The data stream enters Fig. 1 as a series of sections of predetermined length. Preferably the data stream consists of a series of concurrent packets
20 interleaved. For example, there may be up to 8 packets, which are divided into sections (e.g. of 8 bytes at a time), the sections of different data packets being interleaved. The steps of Fig. 1 are performed on one of these sections at any time, using information about which of the packets the section comes from. This means that there is no need to buffer the entire data stream as it
25 arrives.

The first step (step 5) of the operation is to determine the layer 3 start offset, and whether the data is VLAN tagged or SNAPed. To begin with, for each 8 bytes received the algorithm calculates various variables as follows. Firstly, it
30 updates a count variable (*length*) which indicates the number of bytes of the

packet received so far, by adding the number of new bytes to the previous value of *length*. A variable *index* is defined as the largest integer which is no greater than *length* divided by 8. A variable *offset* is then defined as *length* modulus 8.

5

Fig. 3 shows the variables *index* and *offset* for 4 types of packet, labelled (a), (b), (c) and (d) having the data shown in byte locations marked by the row marked as "byte number".

- 10 • Packet type (a) is not VLAN tagged or snapped, and layer 3 starts at byte 14.
- Packet type (b) is VLAN tagged (so that bytes 12 and 13 are 0x8100 (a hex notation) and layer 3 starts at byte 18.
- 15 • Packet type (c) is SNAPped with snap encapsulation starting at byte 14, and layer 3 data starts at byte 22. The bytes at positions 14 to 19 are 0xAA-AA-03-00-00-00.
- Packet type (d) is SNAPped with snap encapsulation starting at byte 18 and also VLAN tagged (so that bytes 12 and 13 are 8100), and layer 3 data starts at byte 26. The bytes at positions 18 to 23 are 0xAA-AA-03-00-00-00.

20

To determine the position of the L3 start, the following steps are performed at a time when the variable *length* is such that *index* is 1:

- 25 • Check the bytes at offset 4 and 5. If they are not 0x8100 and not less than 1500, then the byte is type (a), and layer 3 starts at byte 14.
- Otherwise, if the bytes at offset 4 and 5 are 0x8100, then the packet is tagged (the packet must be type (b) or type (d)). Set a variable *tagged* to be equal to 1.

- Otherwise, if the bytes at offset 4 and 5 are less than or equal to 1500, and the bytes at offsets 6 and 7 are 0xAA-AA, then the packet must be snapped. Set a variable *snapped* to be equal to 1.
- Otherwise, the packet is in an unknown protocol.

5

When the next section of the data packet arrives, so that the variable *length* is such that index is 2:

- Check the bytes at offsets 0 and 1. If they are greater than 1500 and *tagged*=1, then the packet is type (b) and layer 3 begins at byte 18.
- 10 • Otherwise, if the bytes at offsets 0 and 1 are less than or equal to 1500 and *tagged*=1, then set *snapped*=1.
- If *tagged*=1 and *snapped*=1 and the bytes at offsets 2 to 7 are AA-AA-03-00-00-00, then the packet is type (d), and layer 3 starts at byte 22.
- Otherwise, if *tagged*=0 and *snapped*=1 and the bytes at offsets 0 to 3
15 are 0x03-00-00-00, then the packet is type (c), and layer 3 starts at byte 22.
- Otherwise, the protocol is unknown.

Referring once more to Fig. 1, once the positions of the start of the layer 3
20 (and other layers) are known, the section of the data stream is passed to a first parser 7 and to a second parser 9 as discussed below. Note that the step 5 operation, and the first parser 7 and second parser 9 operations are performed on one of these sections at any time. In this case, the step 5 operation uses section identity information identifying which packets the
25 section of data belongs to, and for example in the case that there are multiple packets maintains a set of variables (e.g. variable *length*) for each of those packets. In the processing of a section of the data stream derived from a given packet, step 5 involves updating the variables for the corresponding packet. The parsers 1 and 2 do not have to know this information however.

30

The first parser 7 extracts data from the packet according to positions defined by a set of registers 8. For example, when 8 bytes are to be extracted, 8 registers (labelled Offset reg 1, ..., Offset reg 8) are used. Each register holds an indication ("L2/L3") of whether data is to be extracted from the layer 2 or
5 laer 3 data, and also an offset indicating which bytes are to be extracted relative to this starting positions of those layers. In this way the first parser 7 is able to extract local keys. The extracted local keys are compared in an AND operation 11 with 8 16-bit masks 15 (each of the 8 registers extracts 16 bits). The same 8 16-bit masks 15 are compared with 8 16-bit rules 17 by an AND
10 operation 13. The results of the AND operations 13 and 11 are compared in step 19 to produce 8 1-bit results.

Meanwhile the second parser 9 receives the same data stream and the results of the determination of the start of the layers, and extracts a set of 8
15 bits determined by 8 key selection registers 23. The outputs of the second parser 9 are compared with those of the compare operation 19 in a step 21.

The operation of the second parser 9 and of the combine unit 21 is shown in Fig. 2. The upper portion of Fig. 2 shows the conventional structure of a data packet, starting with layer 2 data ("L2 info"), then layer 3 data ("L3 info"), then
20 layer 4 data ("L4 info"). Using the results of the layer position determination algorithm, the second parser 9 is fed selected ones of the bytes as shown of Fig. 2. According to the outputs of programmable selector 23, the MUX multiplex units 25 output one of their inputs. These are fed to further MUX
25 multiplex units 27, 29. The MUX units 27 receive other portions of the data packet, and also output of the first parser 7. The MUX units 29 receive the respective outputs of the MUX units 27, and also of the respective MUX units 25. MUX units 27, 29 are controlled based on selection signals sel[0] 31, sec[3] 31, which also come from the programmable registers 23. The result is
30 the extracted key for the flow engine.

Referring to Fig. 1 again, a combination 21 of the outputs of the compare operation 19 and the key for the flow engine is made, to generate a final key. The uses of this key will be clear to a skilled reader, as will the exact
5 operation of the two parsers.

Fig. 4 shows the layout of a parser system circuit 35 for implementing the steps of Fig. 1 in the context of an Ethernet switch. The parser system circuit 35 operates on 8 bytes at a time, and has an input interface 41 which receives
10 inputs from a buffer rx_ififo 39 which receives packets from the pins of the Ethernet switch, and also from a MAC interface rx_max_ififo 37 which provides control information including an index identifying the packet description associated with the corresponding packet (this constitutes the section identification information discussed above). The step 5 operation of
15 Fig. 1 is implemented by a unit 43, and the results transmitted *inter alia* to a unit 45 which functions as the first and second parser and performs the combinations shown in Fig. 1 to generate the final key. The unit 45 receives the other data it requires, such as the data of registers 9, 17 and 15 of Fig. 2, from a register file 47. The parser puts all the information for each stream of
20 data (i.e. for each of the concurrent packets) within the packet descriptor for the corresponding packet. By operating on 8 bytes at a time with 2 cycles per processing step the parser is able to manage 8 FE streams.

The output of the unit 45 passes to an output interface 49 of the parser
25 parser_mem_iface, which in turn passes it to other components of the Ethernet switch, in particular to a memory manager rx_mem_mgr 51. Note that all the circuitry of the parser system circuit 35 is preferably implemented on a single integrated circuit.

Claims

1. A parser system having:

an interface to receive a data stream composed of interleaved sections of a plurality of different packets, and section identity information about each
5 of the sections of data defining which packet it relates to,

parsing means for processing the data stream section-by-section and employing the section identify information to identify and extract data.

2. A parsing system according to claim 1 further including user-programmable registers for storing offset information, the parsing means
10 being arranged to identify structural features of the packets using the section identity information and the data sections themselves, and to employ offset data stored in the registers to identify and extract data from the packets in locations defined by the structural features of the packets and the offset data.

3. A parser system having:

15 an interface to receive a data stream composed of a series of data sections which are sections of packets;

one or more user-programmable registers,

parsing means for receiving the sections sequentially and employing offset information stored in the registers to identify and extract data from
20 them.

4. A parser system according to claim 1, claim 2 or claim 3 in which the parsing means includes a scanning section for obtaining structural data identifying the location of layers of data in the packets, and a parser section which uses the output of the scanning section and offset information to extract
25 the data.

5. A parser system according to claim 4 when dependent on claim 2 or claim 3 in which the offset information for at least part of the data to be extracted is the offset information stored in the user-programmable registers.
6. A parser system according to claim 5 in which the parser means
5 comprises a first parser which extracts data identified using offset information stored in the user-programmable registers, and a second parser which extracts data using predetermined offset information.
7. A parser system according to any of claims 4 to 6 in which the scanning system is adapted to identify tags in the data packets.
- 10 8. A method of parsing a data stream comprising transmitting it as a series of sections to a parsing system according to any preceding claim, the parsing system processing it section-by-section.

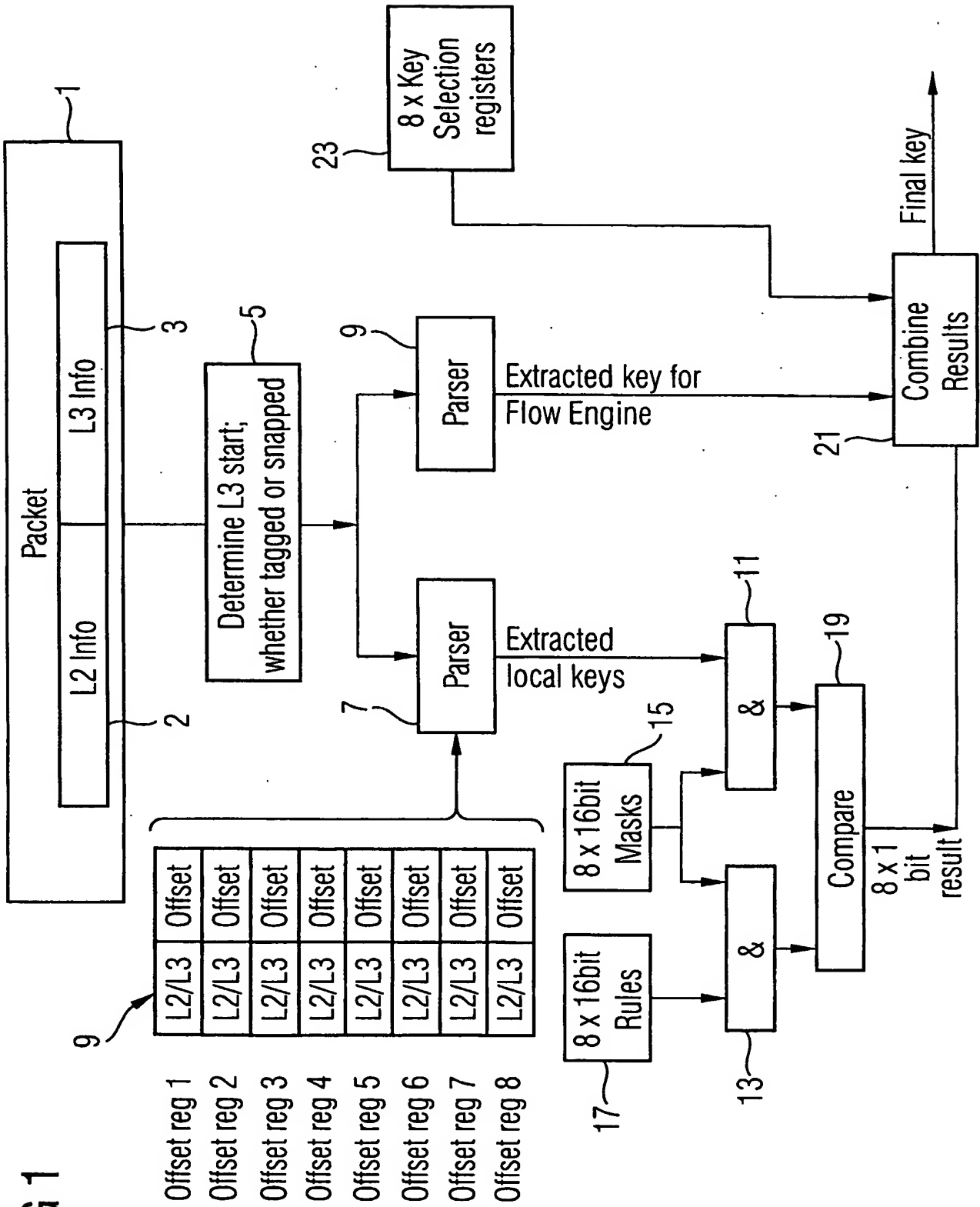


FIG 2

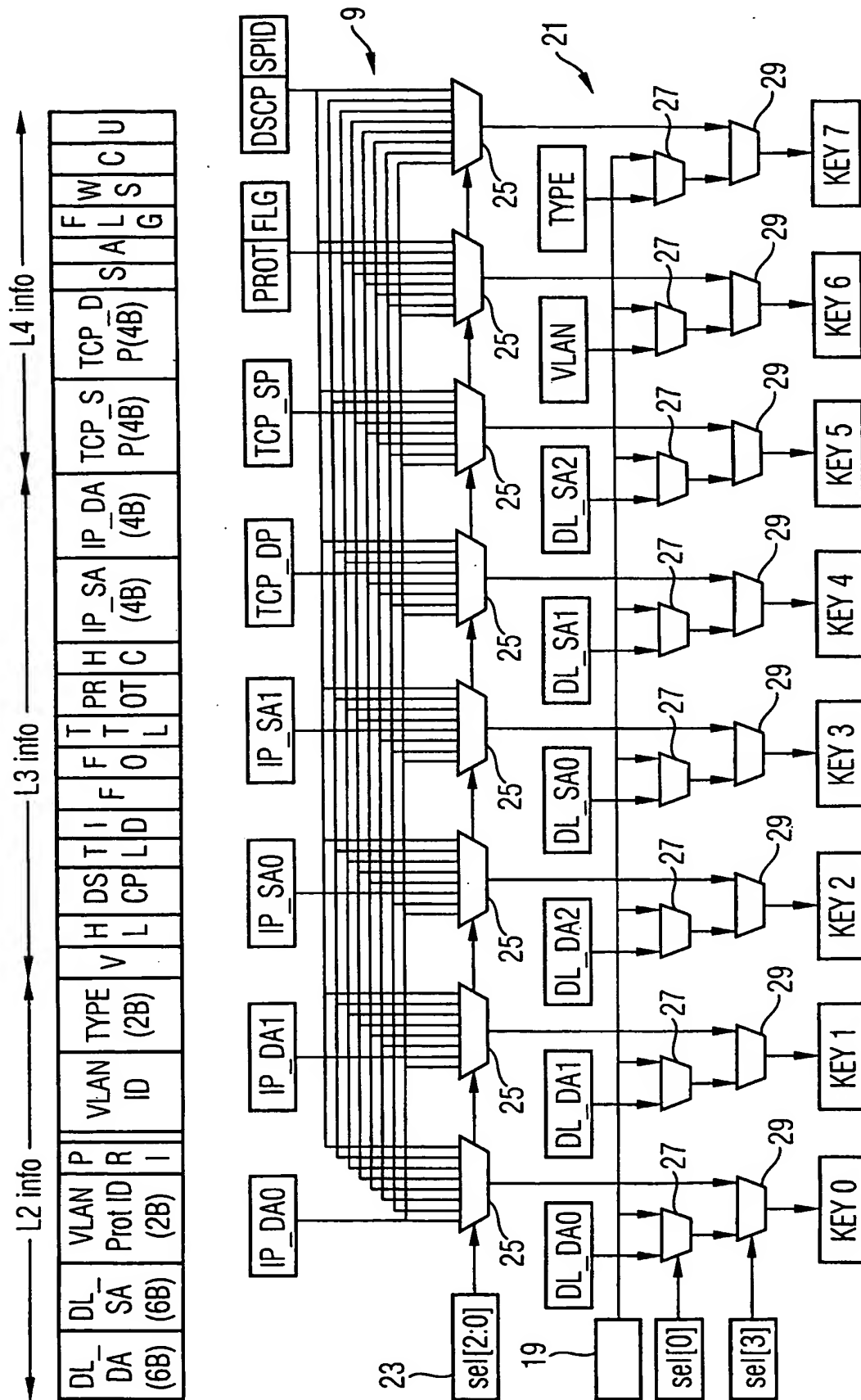
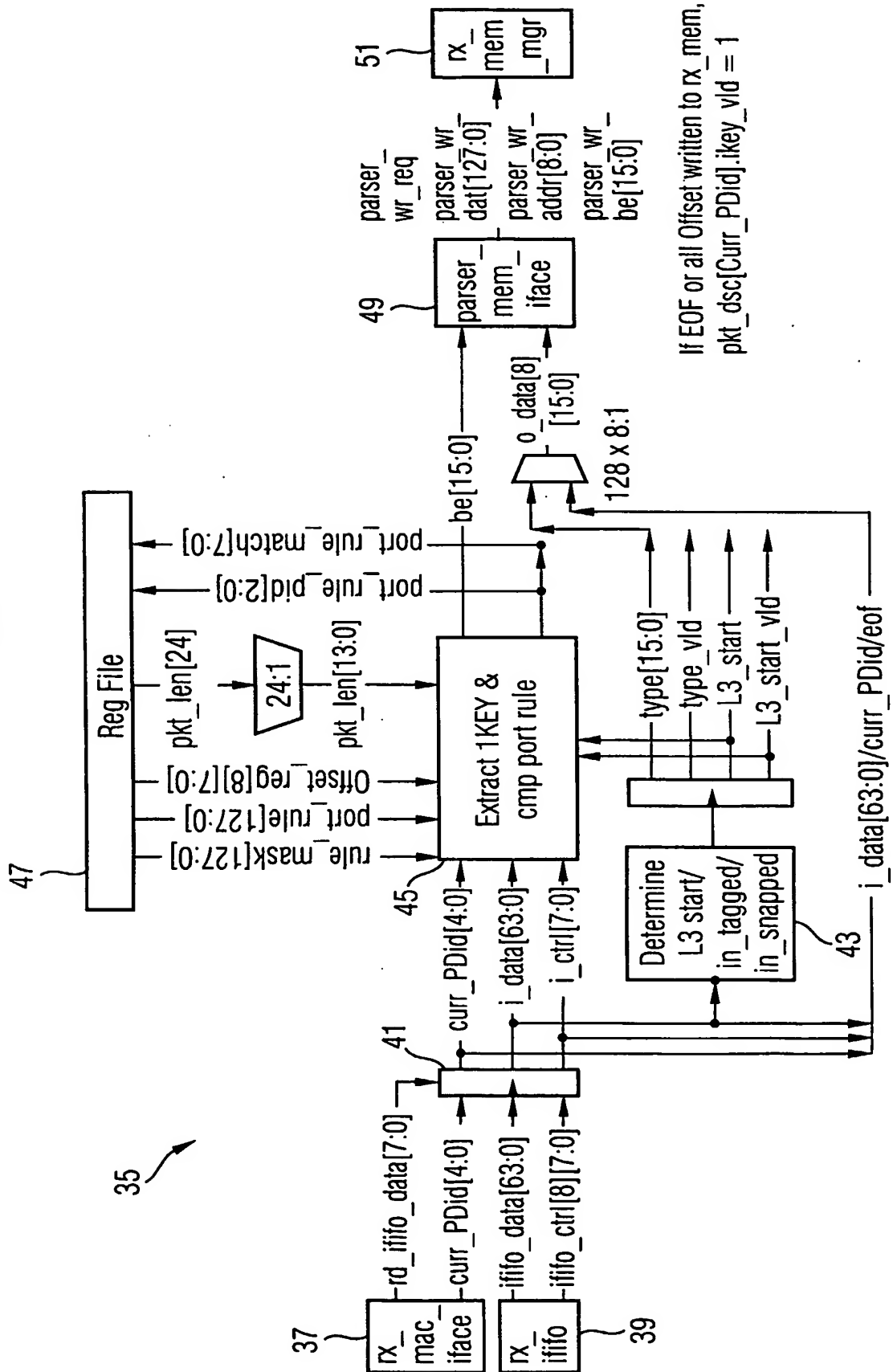


FIG 4



INTERNATIONAL SEARCH REPORT

International Application No.

PCT/SG 02/00210

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 H04L29/06 H04L12/46

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 7 H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 01 33774 A (ADVANCED MICRO DEVICES INC) 10 May 2001 (2001-05-10) abstract page 2, line 18 -page 4, line 10 page 6, line 18 -page 9, line 25 page 10, line 28 - line 33; figures 1-7 --- -/--	1-8



Further documents are listed in the continuation of box C.



Patent family members are listed in annex.

* Special categories of cited documents:

- *A* document defining the general state of the art which is not considered to be of particular relevance
- *E* earlier document but published on or after the international filing date
- *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- *O* document referring to an oral disclosure, use, exhibition or other means
- *P* document published prior to the international filing date but later than the priority date claimed

T later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

X document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

Y document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

& document member of the same patent family

Date of the actual completion of the international search

16 Apr11 2003

Date of mailing of the international search report

25/04/2003

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Jimenez Hernandez, P

INTERNATIONAL SEARCH REPORT

International Application No.

PCT/SG 02/00210

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	<p>WO 00 52897 A (SUN MICROSYSTEMS INC) 8 September 2000 (2000-09-08) abstract page 4, line 19 -page 5, line 11 page 7, line 28 -page 8, line 17 page 9, line 19 -page 10, line 11 page 10, line 31 -page 11, line 26 page 15, line 23 -page 16, line 20 page 17, line 25 -page 18, line 15 page 23, line 7 -page 24, line 27 page 29, line 28 -page 30, line 5 page 30, line 26 - line 33 page 39, line 3 -page 40, line 23 page 52, line 12 - line 19; figures 1-7</p>	1-8
X	<p>US 5 748 905 A (HAUSER STEPHEN A ET AL) 5 May 1998 (1998-05-05) abstract column 1, line 65 -column 3, line 34; figures 1-12</p>	1-4,8
X	<p>US 5 917 821 A (BEHKI NUTAN ET AL) 29 June 1999 (1999-06-29) abstract column 1, line 45 -column 2, line 60; figures 1-16</p>	1-4,8
A	<p>NIRAJ SHAH: "Understanding Network Processors" XP002208129 Retrieved from the Internet: <URL: http://www-cad.eecs.berkeley.edu/{niraj/papers/UnderstandingNPs.pdf}> 'retrieved on 2002-07-31! the whole document</p>	1-8

INTERNATIONAL SEARCH REPORT

International Application No.

PCT/SG 02/00210

Patent document cited in search report		Publication date	Patent family member(s)	Publication date
WO 0133774	A	10-05-2001	EP 1224773 A1 WO 0133774 A1	24-07-2002 10-05-2001
WO 0052897	A	08-09-2000	US 6356951 B1 AU 3613200 A EP 1159814 A2 JP 2002538731 A WO 0052897 A2	12-03-2002 21-09-2000 05-12-2001 12-11-2002 08-09-2000
US 5748905	A	05-05-1998	AU 4230097 A JP 2001500680 T WO 9809223 A1	19-03-1998 16-01-2001 05-03-1998
US 5917821	A	29-06-1999	AU 703464 B2 AU 1270095 A DE 69425757 D1 DE 69425757 T2 EP 0736236 A1 JP 9511105 T CA 2179613 A1 WO 9518497 A1	25-03-1999 17-07-1995 05-10-2000 19-04-2001 09-10-1996 04-11-1997 06-07-1995 06-07-1995